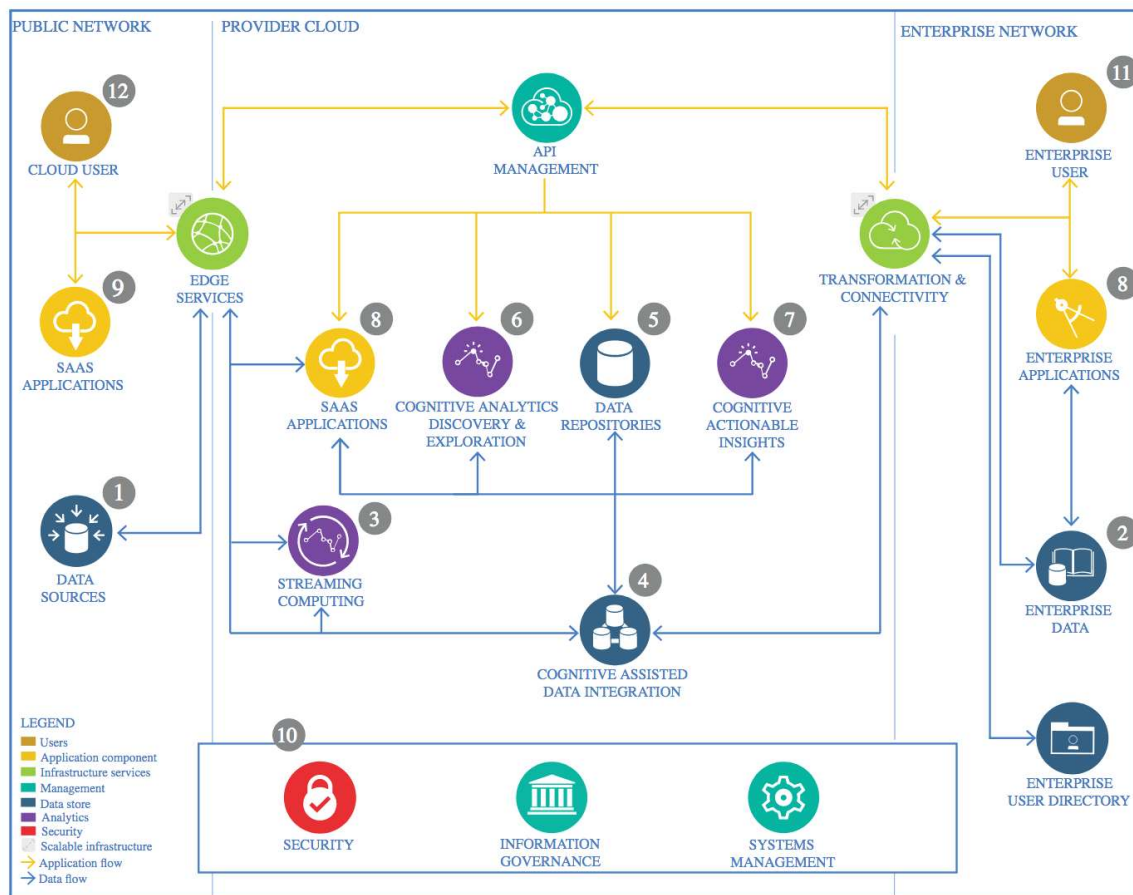


The Lightweight IBM Cloud Garage Method for Data Science

Architectural Decisions Document for Advanced Data Science Capstone

- Architectural Components Overview



IBM Data and Analytics Reference Architecture. Source: IBM Corporation

- Data Source

- Technology Choice

Kaggle Breast Cancer Wisconsin Data Set (<https://www.kaggle.com/uciml/breast-cancer-wisconsin-data>)

- Justification

Kaggle has many data sets in place that are set up for data analysis. Also, I noticed that this dataset would likely work well with XGBoost and a deep learning model due to the variables being exclusively of the same type.

- Enterprise Data

- Technology Choice

N/A

- Justification

N/A

- Streaming analytics

- Technology Choice

N/A

- Justification

N/A

- Data Integration

- Technology Choice

IBM Watson

- Justification

IBM Watson (the lite version) is free, and has many extremely useful features like uploading data and inserting the data as a dataframe. This makes it easier data scientists to spend less time importing their data and more time making improvements

to their model and analysis.

- Data Repository

- Technology Choice

IBM Cloud Storage

- Justification

IBM Cloud Storage is free and has enough storage for the purposes of this project. It is also the default storage on IBM Watson.

- Discovery and Exploration

- Technology Choice

Python Libraries:

Pandas, Matplotlib

- Justification

Pandas dataframes are the gold standard in data science. Easy to manipulate with a very user friendly library and online user support. Matplotlib was used for its easy to create data visualisations.

- Actionable Insights

- Technology Choice

XGBoost

Keras

Tensorflow

Matplotlib

- Justification

XGBoost was used because it is known to be an extremely fast and efficient model to use for classification. XGBoost is the model of choice for many Kaggle data science competitors. Also, it is very easy to implement.

Keras was used because it uses high level code, allowing us to run tensorflow (lower level). This makes us able to run some very complex tasks with just a few lines of code.

- Applications / Data Products

- Technology Choice

PowerPoint Report

- Justification

Powerpoint allows for a very structured presentation and, although boring when used incorrectly, is still a useful way of presenting information.

- Security, Information Governance and Systems Management

- Technology Choice

N/A

- Justification

N/A